
Protocols in Pharmaceutical Sciences

Quantifying Trace Amounts of Aggregates in Biopharmaceuticals Using Analytical Ultracentrifugation Sedimentation Velocity: Bayesian Analyses and *F* Statistics

Lucas Wafer,^{1,2} Marek Kloczewiak,¹ and Yin Luo¹

Received 15 March 2016; accepted 20 April 2016; published online 16 May 2016

Abstract. Analytical ultracentrifugation–sedimentation velocity (AUC-SV) is often used to quantify high molar mass species (HMMS) present in biopharmaceuticals. Although these species are often present in trace quantities, they have received significant attention due to their potential immunogenicity. Commonly, AUC-SV data is analyzed as a diffusion-corrected, sedimentation coefficient distribution, or *c*(*s*), using SEDFIT to numerically solve Lamm-type equations. SEDFIT also utilizes maximum entropy or Tikhonov-Phillips regularization to further allow the user to determine relevant sample information, including the number of species present, their sedimentation coefficients, and their relative abundance. However, this methodology has several, often unstated, limitations, which may impact the final analysis of protein therapeutics. These include regularization-specific effects, artificial “ripple peaks,” and spurious shifts in the sedimentation coefficients. In this investigation, we experimentally verified that an explicit Bayesian approach, as implemented in SEDFIT, can largely correct for these effects. Clear guidelines on how to implement this technique and interpret the resulting data, especially for samples containing micro-heterogeneity (*e.g.*, differential glycosylation), are also provided. In addition, we demonstrated how the Bayesian approach can be combined with *F* statistics to draw more accurate conclusions and rigorously exclude artifactual peaks. Numerous examples with an antibody and an antibody-drug conjugate were used to illustrate the strengths and drawbacks of each technique.

KEY WORDS: analytical ultracentrifugation; bayesian analysis; monoclonal antibody; SEDFIT; sedimentation velocity.

INTRODUCTION

The majority of purified commercial protein pharmaceutical preparations are accompanied by small quantities of product-related impurities, including aggregates of the protein product (1–3). The presence of these aggregates, also referred to as high molar mass species (HMMS)¹ or high molecular weight species (HMWS), has raised numerous safety concerns. This is primarily due to the potential immunogenicity of aggregates larger than dimer (4,5). However, there are also significant difficulties in measuring and predicting various properties of the aggregates, including concentration, oligomerization state, and stability over a drug's shelf life. These concerns are especially pronounced for monoclonal antibodies, which are often delivered in high-concentration formulations of >50 mg/ml (6). The enhanced

potential for self-association has resulted in the aggregate concentration becoming a critical quality parameter during antibody production, purification, and administration.

Size exclusion chromatography–multi-angle light scattering (SEC-MALS), asymmetric flow field-flow fractionation–multi-angle light scattering (AF₄-MALS), and analytical ultracentrifugation–sedimentation velocity (AUC-SV) are currently the most common methods for quantifying low levels of aggregates (2,3,7). However, each of these techniques has its own limitations. For example, SEC-MALS requires a solid-phase separation matrix, which may interact with the protein. In addition, it requires sample dilution, provides limited resolution, and may exclude large aggregates through a “sieving effect” (8,9). AF₄ offers resolution over a limited size range and is generally recognized as being less robust than SEC-MALS. In addition, the sample is dynamically concentrated and diluted during the separation, which can alter the number and size of aggregates present (7). In contrast, AUC is based on first principles, with no solid-phase separation matrix that could interact with the proteins, and therefore it does not suffer from these same drawbacks (10). Experiments can be performed in the formulation buffer

¹ Analytical Research and Development, Biotherapeutics Pharmaceutical Sciences, Pfizer Inc., 1 Burtt Rd, Andover, Massachusetts, USA.

² To whom correspondence should be addressed. (e-mail: Lucas.Wafer@Pfizer.com;)

developed for the therapeutics, which is often not optimal for SEC and AF₄ due to frequently encountered protein-matrix interactions and excipient-membrane interactions, respectively. AUC also has the advantage of separating protein species over a wider range of hydrodynamic size. In addition, if a refractive index detector is used to collect data, higher protein concentrations can be used, which are more representative of the therapeutic drug product. The primary disadvantage of AUC is the expertise required to prepare samples, perform the experiments, and properly analyze the data (10,11). However, advances over the last 20 years have significantly improved data analysis and spurred wider use of AUC for a number of applications (10,12). These include the characterization of novel pharmaceutical proteins and biosimilars, from early stage characterization and formulation development through stability and late-phase comparability studies (13–19). These applications were made possible by the emergence of highly advanced data-analysis packages, the most versatile of which is SEDFIT (17,18,20).

It is outside the scope of this paper to discuss the mathematical algorithms underlying these programs, which are detailed elsewhere (15,17,19). Briefly, SEDFIT numerically fit the ultracentrifugation data to the Lamm equation, for which an analytical solution does not exist (21). For velocity experiments (*e.g.*, AUC-SV), the outcome of the analysis is commonly expressed in terms of a sedimentation coefficient distribution, *c*(*s*). During the experiment, different species will sediment with specific rates, measured in Svedberg (*S*) units ($1\text{ S} = 1 \times 10^{-13}\text{ s}$), depending on their specific hydrodynamic properties. Although the *c*(*s*) is generally used for the characterization of protein mixtures, including the relative abundance of protein species and the detection of protein-protein interactions, it has several important limitations. For example, the fitted results associated with a *c*(*s*) are sensitive to the initial fitting parameters, since there is not an analytical solution to the Lamm equation. The fitted parameters include the position of the meniscus (22), the type of noise (23), the resolution (24), the confidence level (25,26), and the integration range. The use of regularization is also known to lead to small but systemic errors in the sedimentation coefficients and relative abundance, especially for species at trace levels (26). In addition, several investigators have reported variable results depending on the condition of the hardware, which includes the type and age of cell centerpieces, windows, and housing (27); the alignment of cells in the rotor (27); the rotor temperature control (23,28); and the condition of the rotor itself (16). Nevertheless, if AUC-SV is performed with due diligence, trace amounts of HMMS can often be detected and quantified.

The aim of this work is to re-investigate our ability to use AUC-SV for detecting and quantifying HMMS in biopharmaceuticals using the advance tools available in SEDFIT, including both the explicit Bayesian model and the *F* statistics calculator. Specifically, we address if, and when, trace oligomeric species present in the normal *c*(*s*) are significant and necessary to fit the data. In addition, we experimentally explore the impact of a Bayesian analysis on the sedimentation coefficients and relative abundance of HMMS with a background of a large amount of monomeric protein. The similarities and differences between the automatic and manual Bayesian tools are also demonstrated.

These analyses can help to answer questions such as when the results can be trusted, and how to test different hypotheses regarding the composition of a given sample. To the best of our knowledge, the combination of Bayesian tools and *F* statistics have only been applied to simulated data sets, and an in-depth discussion of the applications of these tools to experimental data has not been reported.

MATERIALS AND METHODS

Materials

IgG-ADC, an antibody-drug conjugate, was made by covalently attaching a small molecule drug to a CHO-expressed monoclonal IgG4κ antibody. This molecule was produced and purified by Pfizer Inc. (New York, NY, USA). Immunoglobulin A (IgA) purified from human colostrum, Tris-HCl, and sodium chloride were purchased from Sigma (St. Louis, MO, USA). Gel filtration Superdex 200 column was purchased from GE Healthcare Life Sciences (Pittsburgh, PA, USA). The cell housings and their components used in the analytical centrifugation experiments were purchased from Beckman Coulter (Brea, CA, USA).

AUC-SV Experiments

Sedimentation velocity experiments were carried out using two Beckman Coulter XL-I analytical centrifuges equipped with absorbance optical systems. Experiments were performed using either one of the two 4-hole An60 Ti rotors or the 8-hole An50 Ti rotor. All cells contained sapphire windows and 12-mm charcoal-filled Epon double-sector centerpieces. The following conditions were used in all sedimentation velocity experiments: 40,000 rpm angular velocity, 20°C rotor temperature, and 280-nm absorbance scanned between 5.8 to 7.3 cm radial distances with radial scanning increment of 0.003 cm. The reference cell contained 420 μl of buffer, and the sample cell contained 410 μl of protein in the same buffer. After reaching 20°C, the rotor was equilibrated for an additional hour before starting the sedimentation run (29). Absorbance data were collected for each experiment for a minimum of 120 scans and a maximum of 300 scans.

AUC-SV Data Analysis: Normal *c*(*s*)

Sedimentation data analysis was performed using SEDFIT program version 14.4f (24). All data was initially analyzed using a continuous distribution, *c*(*s*), with maximum entropy (ME) and Tikhonov-Phillips (TP) regularization. In all cases, identical values were used to initialize the fitting parameters: $S_{\min} = 0$, $S_{\max} = 25$, buffer density $d = 1.00585$, buffer viscosity = 0.01031 P, protein partial specific volume = 0.73, frictional ratio = 1.6, and confidence level = 0.7. However, the reported *S* values were not corrected for buffer density and viscosity, as it was unnecessary for the analysis and did not impact any of the conclusions. The resolution parameter was set to 251, equivalent to an effective resolution of 0.1 S, for both the Run and Fit functions of the program. Values of time-independent noise, meniscus (initial =

6.02 cm), baseline, and frictional ratio were allowed to float during the fit.

At least 100 scans were used to fit any given data set. For smaller sets, where the total scan number was less than 130, every scan was used. For larger data sets, where the number of scans was greater than 200, every other scan was used. The fitting limit for data analysis was set approximately 0.02 cm away from the initial position of the meniscus to avoid any optical disturbances typically observed at the meniscus. The position of the upper limit of data analysis was set to fall in the range of 0.05 to 0.07 cm from the bottom of the cell, where a plateau was still visible. The experimental data was fit to generate a $c(s)$ distribution using the Marquardt-Levenberg global minimization procedure, and the tabulated $c(s)$ distributions by ME and TP regularizations were exported to an Excel spreadsheet for peak integration (see *AUC-SV Peak Integration*).

AUC-SV Data Analysis: Automatic Bayesian $c^P(s)$

Sedimentation data analysis was first performed using SEDFIT, with all parameters as described above, to generate a normal $c(s)$. When indicated, the data was modeled with an automatic Bayesian analysis, $c^P(s)$, using only ME regularization. This was performed using the Ctrl + X shortcut implemented in SEDFIT (24) and is equivalent to using the following menu options: Options→Size-Distribution Options→Prior Knowledge of Discrete Species ^X. This analysis uses the $c(s)$ as an input and aims to automatically identify the major species present in the $c(s)$ and fits each with a delta function (24); *i.e.*, this option informs SEDFIT that the user has prior knowledge that the sample contains only discrete species. Following this operation, a distribution for the sample containing only discrete species at the s values obtained from the $c(s)$ will appear in SEDFIT as a dotted red line. Simultaneously, SEDFIT will display the new fit to the raw data, biased by the assumption of discrete species, as a solid black line. The resulting peaks in the new distribution, $c^P(s)$, were integrated and analyzed as described in *AUC-SV Peak Integration*.

AUC-SV Data Analysis: Manual Bayesian $c^{MP}(s)$

The manual Bayesian analysis, $c^{MP}(s)$, was performed by first generating a normal $c(s)$, and then by using the Ctrl + W shortcut implemented in SEDFIT (24). This is equivalent to using the following menu options: Options→Size-Distribution Options→Use Prior Probabilities ^W. This option informs SEDFIT that the user has some degree of knowledge about the sample, but does not want to incorporate this information as hard constraints in the fitting. Therefore, the program incorporates the user-provided information as “prior probabilities,” in contrast to “prior knowledge” in automatic Bayesian operation. In the prior probability operation, the total number of species used for initialization was manually varied in a systematic manner, from one to four, to assess the impact on the analysis. The peak width was initialized to 0.5 for all species to allow for a minimum degree of heterogeneity within each species, which helps to improve the quality of the fit based on our experiences. The sedimentation coefficients and amplitudes were initialized using values obtained from

the initial $c(s)$. Whenever these initial values for the manual Bayesian were altered, the experimental data was first re-fit to generate a normal $c(s)$. To test the robustness of the manual Bayesian analysis, several inaccurate seed values were also used to initialize the $c^{MP}(s)$. These included artificial sedimentation coefficients, which differed by at least 1 S relative to their expected values, as well as artificial amplitudes, which differed several-fold from their expected values. The resulting fits, $c^{MP}(s)$ distributions, were integrated and analyzed as described in “*AUC-SV Data Analysis: Peak Integration*” section.

AUC-SV Data Analysis: Peak Integration

The calculations of peak areas and weight-average sedimentation coefficients require an analyst to define the peaks of interest. In SEDFIT program, the peak selection is performed graphically *via* moving the mouse within a $c(s)$ profile. This operation carries uncertainties in mouse positions that impede the reproducibility of the results, even for repetitive calculations for the same data. When more robust calculations were desired, the $c(s)$ data were exported to Excel, and the peak area and weight-average sedimentation coefficient were calculated using the following equations:

$$\text{Area} = \sum h \frac{(C(S)_i + C(S)_{i+1})}{2} \quad (1)$$

and

$$S_{\text{avg}} = \frac{\sum (C(S)_i \times S_i^2)}{\sum (C(S)_i \times S_i)} \quad (2)$$

where h is the resolution of sedimentation coefficient and $c(s)$; i is the optical density for given sedimentation coefficient, s_i . The index parameter, i , is counted from the beginning to the end of a peak of interest. Excel was used because it permitted adjustable, numerical peak selection, which allowed for reproducible comparisons between corresponding peaks across several different experiments. In addition, one could examine the data at a later date without having to regenerate the $c(s)$ first by repeating the analysis in SEDFIT. It should be noted that a similar level of functionality could also be achieved by using SEDFIT’s option for “integration ranges from a file.”

AUC-SV Data Analysis: Non-Interacting Discrete Species Model and F Statistics

All AUC-SV data was fit with the non-interacting discrete species model in SEDFIT. This model allows for up to four “ideally sedimenting species” to be simultaneously fit to the data (24). All samples were initially fit with models containing four discrete species. All species-specific parameters (*e.g.*, c , S , and M) were floated. The concentration parameter, (c) , for component 1 was initialized to the total optical density for a given sample (*e.g.*, 1.0) and for components 2–4, the initial values were relative to that of component 1 (*e.g.*, 5% of the total optical density would be

0.05). The molecular weight parameter, (M), and the sedimentation coefficient parameter, S , were initialized to the corresponding values obtained from the normal $c(s)$ (e.g., ~150 kDa and 5.9 S for component 1). As with the normal $c(s)$, the meniscus and time-independent (TI) noise parameters were floated. For each sample, the critical value of root mean square deviation (RMSD) was determined using the F statistics calculator in SEDFIT by using the following menu options: statistics→Calculate variance ratio (F statistics). The default values were used for the confidence level (0.683), as well as the first and second degrees of freedom. To reduce the likelihood of the fit being trapped in local minima, both Marquardt-Levenberg and Simplex global minimization algorithms were used on each data set until there was no change in the RMSD between two successive fits. Following this, the Marquardt-Levenberg algorithm was run an additional three times to ensure the stability of the RMSD. To determine the significance of the minor species in a sample, this fitting process was systematically repeated using multiple models; each subsequent model differed from the previous by removing the least abundant species. In other words, if the first model included the monomer, dimer, and an HMMS, the second model would only include the monomer and dimer. Individual species were defined to be statistically significant if their removal from the fit resulted in an RMSD larger than the critical RMSD.

RESULTS

Impact of the Regularization Methods

One of the fundamental difficulties in the algorithm to generate a sedimentation coefficient distribution, $c(s)$, is that the process requires an inversion of the Fredholm integral. This is an ill-posed problem because the solution is not unique. There are an infinite number of solutions which describe the data equally well, for any pre-defined threshold for statistical precision. Furthermore, the subset of solutions that most optimally fit the data is dominated by high-frequency oscillations. These often obscure the underlying information and preclude meaningful interpretation of the data, especially for trace components. To address this issue, SEDFIT employs two different types of regularization: maximum entropy (ME) and Tikhonov-Phillips (TP). Both are well-established approaches to minimize oscillations in the solutions to ill-conditioned problems, without significantly impacting the accuracy or precision of the solution.

ME regularization is the default option in SEDFIT (30). This method biases the solutions toward the subset of solutions that contain the highest informational entropy (i.e., the least information), with the implicit assumption that all sedimentation coefficients are, *a priori*, equally likely. This technique has been recommended by Schuck *et al.* for samples containing discrete species (30,31), which is often the case for pharmaceutical preparations of monoclonal antibodies (26,32). TP regularization is also available in SEDFIT, but is based on an alternative set of prior assumptions. This method biases the solution toward those that minimize the second derivative of the coefficient distribution (i.e., those with the least curvature). This technique is generally recommended for samples that contain

broad or heterogeneous distributions, such as synthetic polymers or solutions containing heat-stressed aggregates (30,31). It should be noted that both techniques select the most parsimonious distribution within a pre-defined confidence level and both are valid. However, differences can arise in the $c(s)$ profile depending on the regularization method. In these cases, the assumptions of each approach should be reviewed.

Figure 1 shows representative $c(s)$ distributions using ME and TP regularizations for the same IgG-ADC. Although the integral and sedimentation coefficient of the monomeric species are similar in both cases, there are significant discrepancies in the dimer and HMMS. In this example, there are two additional species present in the sample analyzed with TP regularization. This effect can be seen more clearly in Table I which tabulates the results of ME and TP regularizations on the $c(s)$ distribution. For these data, the results of TP regularization appear to result in a higher abundance of larger HMMS. However, it is also known that the ME regularization tends to systematically under-report HMMS near the detection limit (26). The proper analysis is important, especially because the presence of higher-mass aggregates (i.e., larger sedimentation coefficients), and the relative abundance of aggregates, is thought to be directly related to immunogenicity of therapeutic proteins. The solution lies in the proper incorporation of Bayesian prior probabilities.

Impact of the Automatic Bayesian Analysis

Bayesian tools were incorporated into SEDFIT to explicitly address some of the aforementioned limitations of the normal $c(s)$. These include the bias toward the predominant species when using traditional regularization (see above), as well as the tendency of regularization to generate artificial “ripple peaks” at large s values. In addition, the Bayesian approach provides a more nuanced method to incorporate prior knowledge into the $c(s)$ distribution, rather than relying on the generic, and often unrealistic, assumptions incorporated into traditional regularization.

The simplest form of an explicit Bayesian analysis within SEDFIT can be run automatically after fitting AUC-SV data to a normal $c(s)$ (see “MATERIALS AND METHODS” section). Figure 2 shows the typical result of the automatic Bayesian approach, called a $c^P(s)$, of an IgG-ADC (i.e., same sample as shown in Fig. 1). The new distribution is a result of SEDFIT applying a rational bias to the normal $c(s)$. In other words, the program generates a second $c(s)$ based on the first, with additional weight being given to the sedimentation coefficients of the detected species. The most striking observation is that the choices of regularization no longer appear to impact the analysis, as the two $c^P(s)$ distributions, based on two $c(s)$ distributions generated by different regularization methods, are nearly identical. This effect is more clearly seen in Table II, which tabulates the impact of the automatic Bayesian approach on the number, sedimentation coefficient, and percent abundance of all species. In addition, neither $c^P(s)$ contains apparent HMMS peaks. Rather, the signal originally distributed to these apparent aggregates has been redistributed across the mAb monomer and dimer species, as well as a category SEDFIT

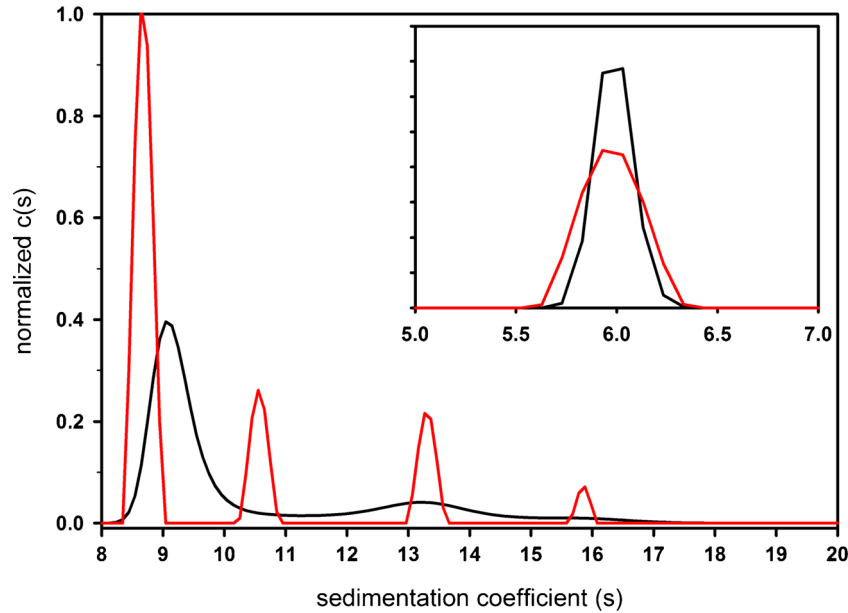


Fig. 1. Overlaid $c(s)$ distribution profiles for the IgG-ADC sample generated using ME (black) or TP (red) regularizations. The sedimentation coefficient ranges are selected to show peaks corresponding to the oligomeric species (main panel) or the monomer (inset). For clarity, the $c(s)$ distribution is arbitrarily scaled relative to the signal of the dimer (main panel) or monomer (inset). The experimental AUC-SV data was obtained with a rotor speed of 40,000 rpm, using absorbance detection with 12-mm path length. The initial sample concentration corresponded to an absorption signal of 0.8 OD_{12 mm}

labels as “OTHER MATERIAL.” The latter does not necessarily refer to any HMMS and the proper explanation of this category is addressed in the following paragraph.

Because the interpretation and implications of the new $c^P(s)$ distribution may be very different than those of the original normal $c(s)$, it is important to emphasize several caveats of this analysis. First, the automatic Bayesian should only be applied to data from samples that are known to contain discrete species. This is because the program forces all identified species to have a peak width of 0, by definition of the delta function, and may therefore produce erroneous results for samples that have broad distributions. Figure 3 shows a normal $c(s)$ distribution generated using ME regularization for a sample containing heterogeneous

aggregates of an IgA antibody, overlaid with its corresponding $c^P(s)$ distribution. The normal $c(s)$ exhibits three broad peaks centered at ~11 S, 13.7 S, and 16.5 S. However, the $c^P(s)$ consists of two sharp, discrete peaks and some broad, artificial peaks before, between, and after the two sharp peaks. This non-typical profile indicates the “prior knowledge,” *i.e.*, the underlying assumption of discrete species applied in the automatic Bayesian, is not applicable to this sample. Specifically, peaks at 10.3 S, and 13 S were generated artificially by SEDFIT to compensate for the additional signal that could not be fit to individual, homogeneous species. It should also be noted that SEDFIT will generate a notepad popup window after the automatic Bayesian is performed, such as that shown in Fig. 3b. Proper

Table I. Comparison of Normal $c(s)$ Distributions for the IgG-ADC Sample Following ME and TP Regularization

	ME		TP	
	S value ^a	% Abundance ^{a,b}	S value ^a	% Abundance ^{a,b}
ADC1				
Monomer	5.98 (0.004)	95.12 (0.37)	5.98 (0.004)	95.30 (0.29)
Dimer	8.87 (0.14)	3.73 (0.47)	8.69 (0.15)	3.18 (0.18)
HMMS 1	12.19 (0.92)	0.77 (0.11)	10.73 (0.17)	0.70 (0.17)
HMMS 2	14.14 ^c	0.11 (0.19)	11.98 ^c	0.21 (0.36)
HMMS 3	n.a.	n.a.	13.10 (0.18)	0.27 (0.31)
HMMS 4	n.a.	n.a.	14.44 (0.37)	0.22 (0.2)

^a The sedimentation coefficient values are the average of triplicate experiments. The standard deviation is listed after the averaged value in parentheses

^b The abundance is expressed in terms of integrated signal in the corresponding s value range, relative to the total signal. The standard deviation is listed after the averaged value in parentheses

^c Observed in a single experiment

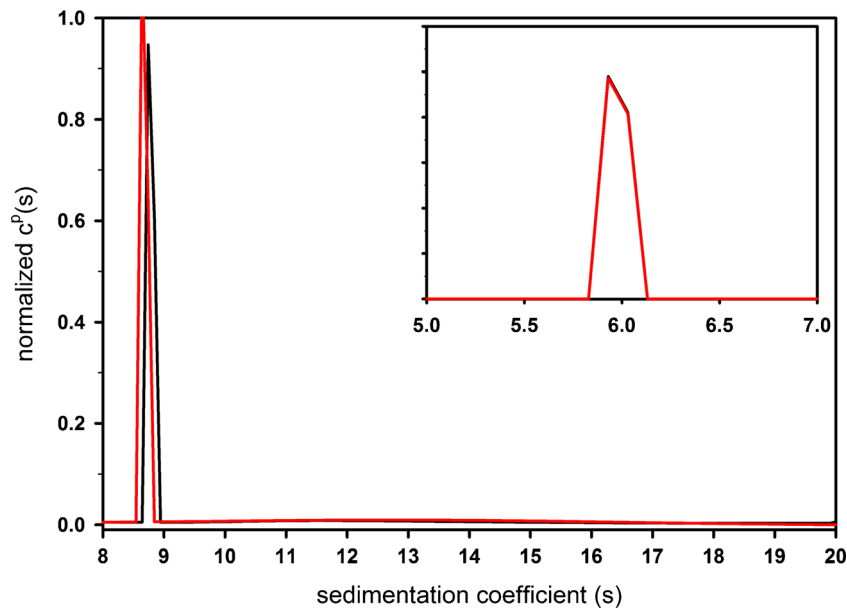


Fig. 2. Overlay of the dimeric (*main panel*) and monomeric (*inset*) peaks in the automatic Bayesian $c^P(s)$ distributions performed following the ME (*black*) or TP regularization (*red*) for the same data in Fig. 1. For clarity, the $c(s)$ scale of the main panel is normalized to the relative abundance of the dimer. The *inset* shows the unscaled monomer peak, which perfectly overlaps for both sets of analyses

care must be taken to correctly interpret this data. For example, the number of peaks listed within the notepad popup may be different than the number detected by the user through visual inspection, or by using the SEDFIT peak detection algorithm Ctrl + M (see “**MATERIALS AND METHODS**” section). In addition, any signal that SEDFIT could not fit to the identified discrete species is placed in a generic category, labeled as “OTHER MATERIAL.” It is our experience that the sedimentation coefficient listed for this category often corresponds to that of another identified peak (*e.g.*, “Peak 2”). However, it is actually a weight-average value that represents the entire unassigned signal. Therefore, it could result from both heterogeneity in the identified species, as well as the presence of small contaminant species, which may exist throughout the distribution but are masked in the $c^P(s)$. Furthermore, analysis using F statistics often demonstrates that the signal assigned to “OTHER MATERIAL” is not significant. Due to these limitations, we recommend that the notepad results should be used for

general guidance and not for the final interpretation of the data.

Impact of the Manual Bayesian Analysis

For cases where the automatic Bayesian is not appropriate for the sample (*e.g.*, heterogeneous samples; see Fig. 3), or additional control of the fitting is desired, a manual Bayesian analysis can be performed. As with the automatic Bayesian, this technique should only be used after the AUC-SV data has already been fit to generate the normal $c(s)$. For this option, the parameter output associated with the $c(s)$ may serve as the “prior probability,” *i.e.*, initial parameters for the subsequent Bayesian analysis. Alternatively, the user may vary the initial parameters to test the sensitivity of the data to different seed values. Figure 4a shows the result of the manual Bayesian approach, or $c^{MP}(s)$, applied to the $c(s)$ of the same IgA sample shown in Fig. 3. Unlike the automatic Bayesian, which failed to properly fit the broad,

Table II. Comparison of Automatic Bayesian $c^P(s)$ Distribution for the IgG-ADC Sample Following ME and TP Regularization

IgG1	ME		TP	
	s value ^a	% Abundance ^{a,b}	s value ^a	% Abundance ^{a,b}
Monomer	5.98 (0.004)	94.77 (0.27)	5.98 (0.004)	94.83 (0.36)
Dimer	8.87 (0.14)	3.60 (0.24)	8.69 (0.15)	3.47 (0.23)
Other material ^c	8.87 (0.14)	1.22 (0.37)	8.70 (0.01)	1.70 (0.13)

^a Reported values are the average of triplicate experiments. The standard deviation is listed after the averaged value in parentheses

^b The integrated signal in the respective integration range relative to the total signal. The standard deviation is listed after the averaged value in parentheses

^c Values obtained from notepad popup and do not necessarily represent HMMS (see text). The standard deviation is listed after the averaged value in parentheses

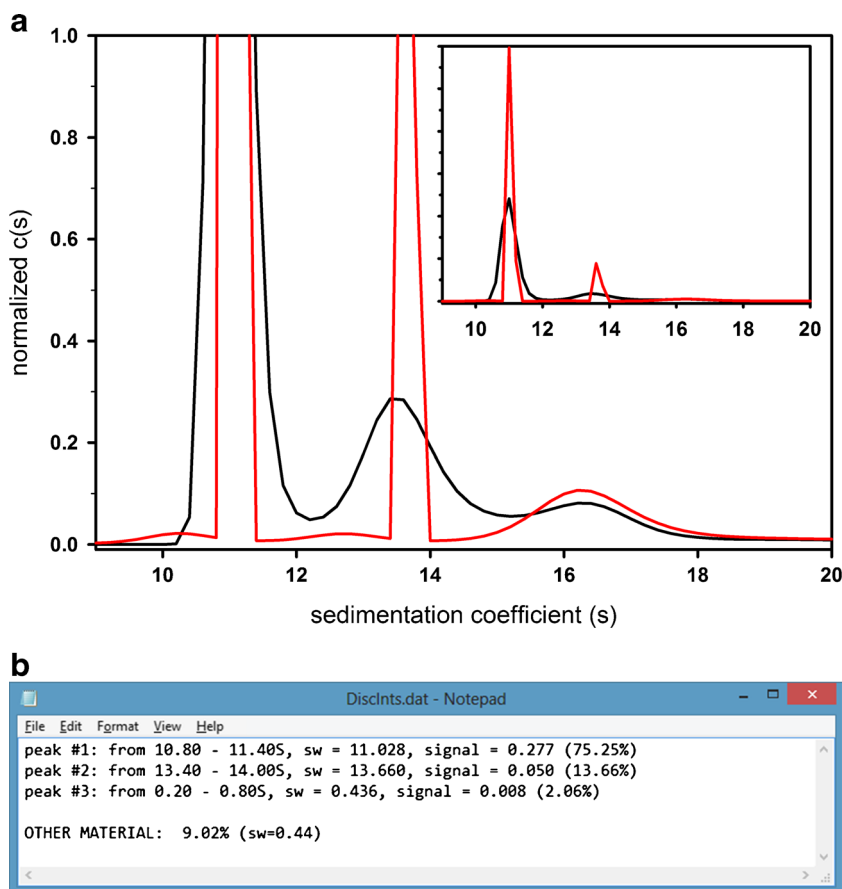


Fig. 3. **a** Overlay of the monomeric and HMMS peaks in a normal $c(s)$ (black) and the corresponding automatic Bayesian $c^P(s)$ (red) for the IgA sample. Both profiles used ME regularization. The AUC-SV data was obtained using the same experimental conditions as Fig. 1. For clarity, the $c(s)$ is shown in the range between 9 and 20 S. The *insert* shows the same $c(s)$ and $c^P(s)$ at the default scale. **b** The notepad popup appeared automatically after the $c^P(s)$ analysis for the same IgA sample

heterogeneous distribution of aggregates, the manual Bayesian shows an excellent agreement between the $c^{MP}(s)$ and the prior probabilities (*i.e.*, predicted distribution based on the results of the initial normal $c(s)$, dashed red line). The primary benefit of this technique is the ability to test the sensitivity of the results to various prior probabilities, rather than simply relying on the general assumptions associated with conventional regularization. For example, when working with recombinant biopharmaceuticals, the user often has information, or reasonable expectations, for the sedimentation coefficients and relative abundance of the monomer and dimer. Using this information, one can determine if the presence of HMMS in the fit depends on specific prior probabilities. This is achieved, for example, by initializing the manual Bayesian analysis with only the monomeric species or changing the relative abundance of monomer and HMMS. The manual Bayesian also allows one to probe the stability of the sedimentation distribution. For example, one can determine to what extent the fit changes in response to artificial, or unreasonable, prior probabilities. Figure 4b shows a representative example of a $c^{MP}(s)$ generated using the same IgA data but a false prior probability. In this example, the sedimentation coefficient

for the dimer peak was initialized at 15 S, rather than the value observed in the original $c(s)$ (~13.7 S). Even though this peak represented a relatively low abundance species, altering the prior probability had a significant impact on the fit. There is a large deviation from the prior probability distribution (dashed red line), which was generated using the initial values (*i.e.*, the *prior probability*—a species at 15 S). In addition, the profile of the distribution has changed, creating a non-symmetrical peak for the affected species and artificially increasing the sedimentation coefficient for the largest species.

The difference observed between the $c^{MP}(s)$ and the prior probability distribution in Fig. 4b provides additional confidence in the existence of the HMMS at ~13.7 S, as well as the sedimentation coefficient determined for this species in the original analysis. The certainty in fitted sedimentation coefficients is especially important for large species in low abundance due to their apparent concentration dependence, as demonstrated in previous work using simulated AUC-SV data and the normal $c(s)$ (26,33). If erroneously increased sedimentation coefficients are not corrected by the manual Bayesian approach, they may mislead investigators to believe in the presence of specific higher-order species. In order to

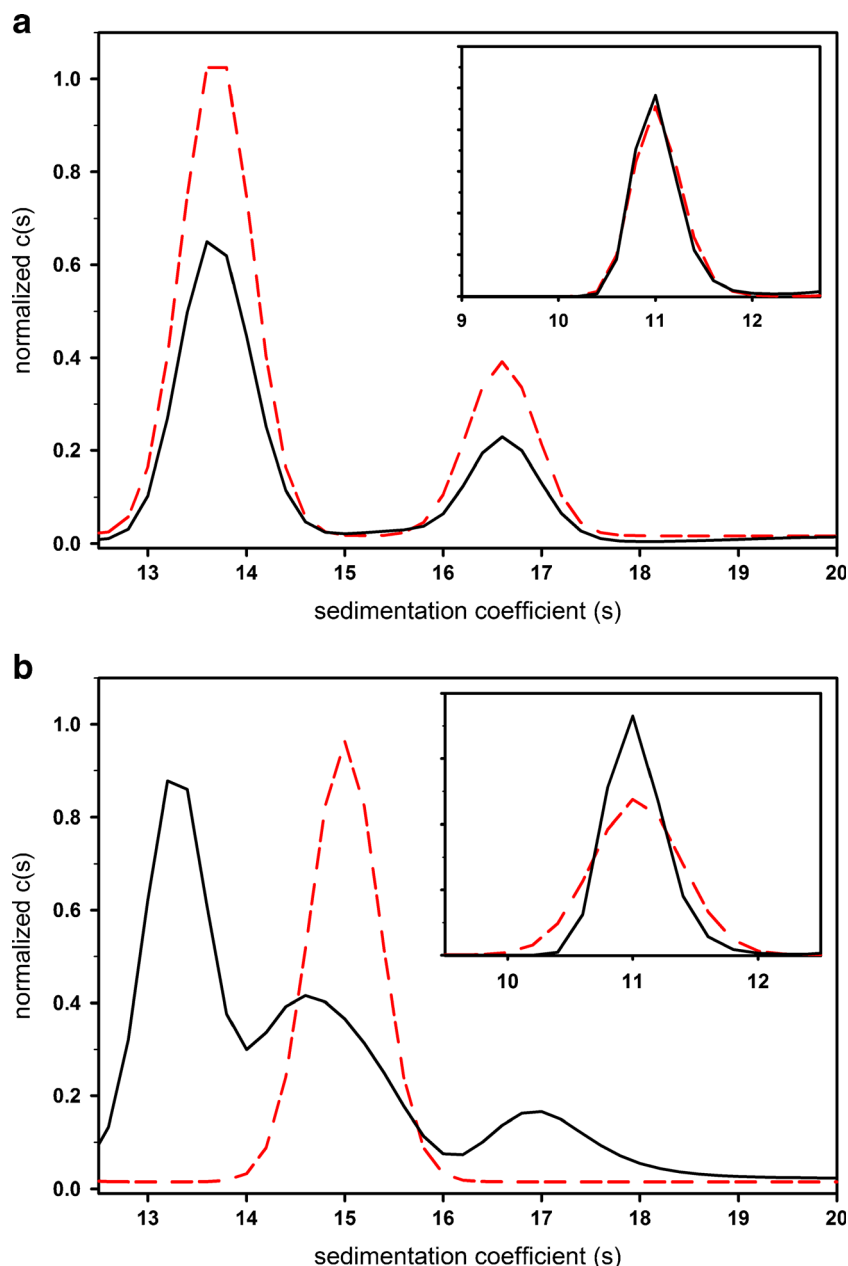


Fig. 4. **a** Overlay of the HMMS (*main panel*) and monomeric (*insert*) peaks in a manual Bayesian $c^{\text{MP}}(s)$ (*black*) and the associated prior probability (*dashed red line*) for the same IgA data used in Fig. 3. The manual Bayesian analysis was performed following an ME regularization. For clarity, the $c^{\text{MP}}(s)$ is shown at an increased scale in the main panel, between 12.5 S and 20 S. The *insert* is independently scaled. **b** Same overlay as in **a**, except that the prior probability (*dashed red line*) included a false sedimentation coefficient for the dimer (see “**MATERIALS AND METHODS**” section)

experimentally verify the observations made with the simulated data, we analyzed the apparent sedimentation coefficients of the IgA monomer and dimer using a wide concentration range.

Figure 5a shows the concentration dependence of the weight-average sedimentation coefficients for IgA monomer (~300 kDa) and dimer (~600 kDa) in normal $c(s)$ generated using ME or TP regularization (filled and open symbols). Both species exhibited a clear trend of higher sedimentation coefficients at lower loading concentrations, in agreement

with the simulated data (26,33). Interestingly, the effect is significantly larger when using TP regularization, as compared to ME regularization. At the lowest concentration tested, the fitted sedimentation coefficient increased by ~35% (18.5 S) relative to the expected sedimentation coefficient (13.7 S). Therefore, the normal $c(s)$ analysis for low-concentration samples would likely lead to erroneous conclusions regarding the stoichiometries, hydrodynamic radii, and/or conformational states of the dimer and all HMMS present in the sample. It is important to note that the artifactual

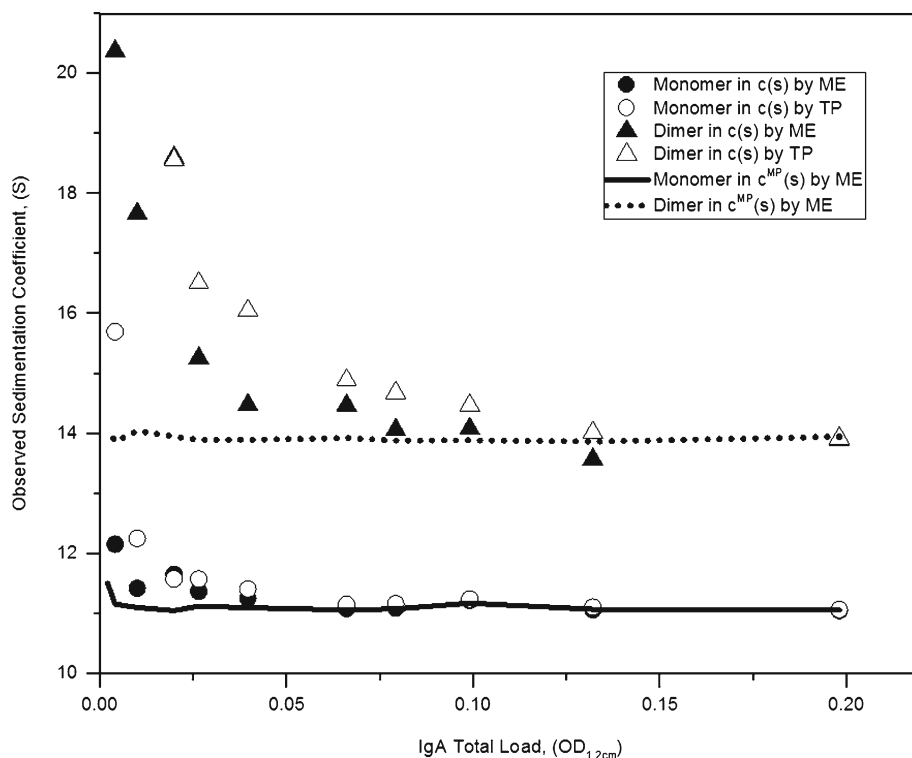


Fig. 5. Concentration dependence of the sedimentation coefficients, S , for IgA monomer (circles) and dimer (triangles) obtained from the normal $c(s)$ with ME (filled symbols) or TP (open symbols) regularizations. Refined values were obtained by performing the manual Bayesian, $c^{MP}(s)$, using the values from the normal $c(s)$ by the ME regularization as the prior probabilities

increase in sedimentation coefficient was observed experimentally at concentrations tenfold higher than those predicted by the simulations ($>0.04 \text{ OD}_{12 \text{ mm}}$ vs. $0.004 \text{ OD}_{12 \text{ mm}}$ or 0.4%, assuming a total load of $1.0 \text{ OD}_{12 \text{ mm}}$). Furthermore, this effect was observed for all HMMS components of IgA (larger HMMS data not shown), which was not predicted by the simulations.

Figure 5 also shows the impact of performing a manual Bayesian analysis on the value of sedimentation coefficients. The solid and the dotted lines indicated that initializing the sedimentation coefficients with the values obtained in the preceding normal $c(s)$ fit (*i.e.*, generating prior probabilities) substantially reduced the spurious increase of the sedimentation coefficients for the monomer and dimer. At the lowest loading concentrations, the apparent shift in the sedimentation coefficient observed for the IgA dimer was $\sim 2\%$ in the $c^{MP}(s)$, as opposed to $\sim 35\%$ in the normal $c(s)$. Similarly, the $c^{MP}(s)$ exhibited less than a 5% shift in the apparent sedimentation coefficient for the IgA monomer, whereas the shift in the same peak in the normal $c(s)$ is $\sim 10\%$.

Determining Significance of HMMS Peaks: F Statistics and Non-Interacting Discrete Species Model

Although the Bayesian analyses offer numerous advantages over the normal $c(s)$, some fundamental questions in biopharmaceutical application may remain unresolved. For example, whether a “ripple peak” with a large sedimentation coefficient represents an aggregate species, or is simply a

numerical artifact. For samples containing discrete species, this issue can be addressed using the non-interacting discrete species model in SEDFIT and the F statistics calculator. The proper applications of these tools allows the user to confidently determine which species are necessary to properly fit the data, and are therefore likely to represent real species in the sample, within a pre-defined confidence interval.

Table III shows representative RMSD values of fitting IgG-ADC data using the non-interacting discrete species model, with the step-wise removal of, first, HMMS, and then, both HMMS and dimer. The resulting RMSD values were compared with the critical RMSD, which was calculated using the F statistics calculator with the assumption of three discrete species (monomer, dimer, and HMMS) to determine the statistical significance of the HMMS and the dimer (see ‘**MATERIALS AND METHODS**’ section). In principle, the removal of any component decreases the degrees of freedom, and therefore should increase the RMSD of the fit. However, when the data was fit with models consisting of only the monomer and dimer species, the RMSD values were below the critical RMSD threshold, indicating that the AUC-SV data can be described equally well with or without the inclusion of the larger HMMS species, within the pre-defined statistical threshold. Therefore, the HMMS resolved in the original $c(s)$ is not statistically significant. On the other hand, when the dimer and HMMS were removed from the fit, the RMSD values of the fits were all larger than the critical RMSD, indicating that the dimer is statistically significant and should be included in the final results of the analysis.

Table III. Determination of Statistical Significance for Species Resolved in the IgG-ADC Sample Using the *F* Statistics Calculator and the Non-Interacting Discrete Species Model

IgG-ADC	RMSD			
	Critical value ^a	Monomer + Dimer + HMMS ^b	Monomer + Dimer ^b	Monomer ^b
Cell 1	0.006805	0.006790	0.006803	0.007903
Cell 2	0.007001	0.006983	0.006996	0.008165
Cell 3	0.006889	0.006871	0.006878	0.007879

^a Calculated using the default confidence level and degrees of freedom for a model including the monomer, dimer, and HMMS

^b Values correspond to the RMSD for a model using only the listed species

As shown in previous sections, there are several caveats associated with the application of this technique that may impact the final conclusion. The most important among these is that the discrete species model can only be applied to samples that consist of non-interacting, discrete species. A concentration series of AUC-SV experiments, or other orthogonal techniques, can help to detect dynamic self-association in a sample, in which case the discrete species model should not be applied. In addition, one needs to consider the total number of potentially significant, discrete species present in the sample. SEDFIT limits the total number of species to four. If more are required, an alternative data analysis package, such as SEDPHAT (20), should be used. Furthermore, the non-interacting discrete species model is sensitive to the floated parameters for each species: concentration, mass, and sedimentation coefficient. It is recommended that these be allowed to float in an unconstrained manner. However, this may occasionally result in spurious values, such as mass values that differ from the expected values by several-fold (*e.g.*, dimer with mass of a pentamer). This can often be addressed by refitting the data with a different set of initial values. If the user instead chooses to constrain a certain parameter, it is critical that the entire analysis, including the calculation of the critical RMSD, be performed with the same constraints. Finally, a valid comparison of the RMSD values requires that the RMSD of each fit has reached its global minimum. There is no explicit method to ensure that a global minimum is reached. Users should rigorously fit the data by, *e.g.*, varying the initial values, to prevent the analysis from being trapped in local minima (see “**MATERIALS AND METHODS**” section; *AUC-SV Data Analysis*).

DISCUSSION

Within the last decade, analytical ultracentrifugation sedimentation velocity has often been used for characterizing oligomeric proteins, often in low abundance, in biopharmaceuticals during their developments. With the proper sample preparation and data acquisition, users can reliably obtain signal:noise ratios of 1000:1 and limits of detection below the RMSD (34). Furthermore, the certainty in data fitting and the applicability of the incorporated models have been substantially improved over the last decade. Surprisingly, these advanced tools to improve the fitting are absent in numerous experimental studies, especially those that define limits of detection and quantitation for AUC-SV

(27,28,33). Consequently, the theoretical gains in accuracy and precision have not been broadly realized.

In the present work, we provided experimental evidence for the effectiveness of the Bayesian approach in analyzing AUC-SV data for biopharmaceuticals. To obtain a *c(s)* profile, the software is required to deconvolute the diffusional broadening that occurs along the monomer-dimer boundary during sedimentation, which is an ill-posed problem (35). Although ME and TP regularization can address this issue to a certain extent, their application requires assumptions that are demonstrably false: ME assumes that the probability for all sedimentation coefficients is equally likely, and TP assumes that the solution with the least curvature is correct. These assumptions can generate problems that are not obvious to users, such as the underestimate of dimeric species and HMMS. In addition, they can lead to artificial increases in the values of the sedimentation coefficients for trace oligomers (26,34). Therefore, even for solutions consisting of three species, the Bayesian approach offers a benefit in terms of both the accuracy of the sedimentation coefficients and the quantitation of HMMS. The use of an explicit (manual or automatic) Bayesian also allows the incorporation of sample-specific information, which is a significant advantage over the normal *c(s)*. For example, one can test whether or not specific HMMS are necessary to fit the data, by initializing with only the monomer and/or dimer in setting the prior probabilities. The result may impact the interpretation of the experimental data. Similarly, one can test specific hypotheses regarding the hydrodynamic properties of individual species, including their precise sedimentation coefficient or relative heterogeneity, which is simply not possible by performing the normal *c(s)* alone.

This work also demonstrates the application of the *F* statistics calculator and the non-interacting discrete species model to experimental data. For well-behaved samples, these tools are invaluable for discerning which minor peak(s), if any, are necessary to fit the data. This may have significant implications for therapeutic antibodies because previously reported LOD and LOQ were based on analyses using only the normal *c(s)*. Furthermore, the majority of these studies used spiked samples consisting of heterogeneous, heat-stressed aggregates (2,36,37). Such samples are not necessarily representative of process-induced aggregation and may not be amenable to proper deconvolution using standard regularization methods, as the heated samples tend to be highly heterogeneous and exhibit non-discrete profiles.

In order to properly address the quantitation of HMMS, we propose a workflow that combines the advantages of the Bayesian

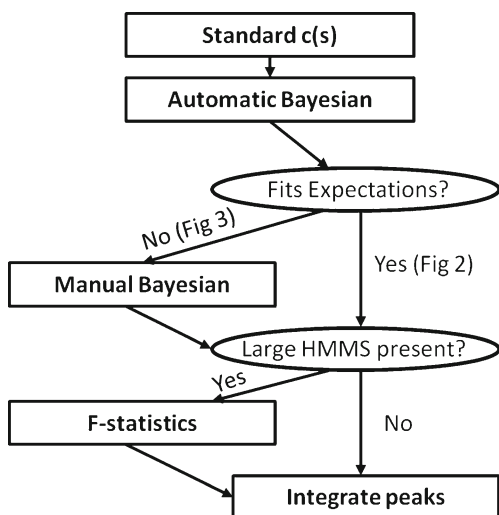


Fig. 6. Flowchart to demonstrate how to use Bayesian analyses and the F statistics tool to enhance AUC analyses, especially for biopharmaceutical samples

approaches and F statistics (Fig. 6). Begin with a normal $c(s)$ and perform the automatic Bayesian in the second step to quickly determine whether the sample can be fit with a model of discrete species (e.g., Fig. 2: IgG-ADC) or whether it requires a model with non-discrete species (e.g., Fig. 3: IgA). If the automatic Bayesian analysis indicates the sample contains only discrete species, the statistical significance of any high s value peaks can be precisely evaluated using the F statistics calculator and the non-interacting discrete species model (e.g., Table III). However, if the automatic Bayesian analysis indicates the sample is heterogeneous with non-discrete species, the manual Bayesian approach should be utilized. This procedure allows specific hypotheses regarding the HMMS to be explored; Fig. 4 provides an example.

Once the final number of species has been determined, the user can manually integrate the peaks present in the $c^P(s)$ or $c^{MP}(s)$ for the analyzed sample. We believe this approach addresses numerous deficiencies in current practices for detecting, quantifying, and characterizing oligomers in low abundance. The application of this integrated approach is especially beneficial for therapeutic antibodies and antibody-drug conjugates. Finally, it should be noted that the explicit Bayesian techniques and the F statistics implemented in SEDFIT cannot overcome the contribution of experimental error to the LOD or LOQ of AUC-SV (2,33,37). These approaches can only reduce the error from the numerical treatment in SEDFIT.

CONCLUSIONS

In the present work, we systematically evaluated the benefits of performing automatic and manual Bayesian analyses, as well as F statistics, following the normal $c(s)$ analysis. We confirmed, with experimental data, that the combination of the automatic and manual Bayesian analyses are powerful tools to determine if a sample is significantly heterogeneous or contains discrete species. Furthermore, we confirmed that these tools correct for the artificial increase in sedimentation coefficients of low abundance species, which has been observed in the normal $c(s)$. For samples consisting

of multiple discrete species in low abundance, F statistics can be applied to rigorously determine their significance. Applying these tools in the correct manner can significantly improve our capability to detect and quantify aggregates in biopharmaceuticals using AUC-SV. A coherent strategy for such an application is demonstrated.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Peter Schuck for his work in developing the SEDFIT software and his education lectures related to its use, Lucy Liu for her AUC analysis of the IgG-ADC, and Qin Zou for his helpful comments and suggestions. This work was supported by Pfizer Inc.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

REFERENCES

1. Filipe V, Poole R, Oladunjoye O, Braeckmans K, Jiskoot W. Detection and characterization of subvisible aggregates of monoclonal IgG in serum. *Pharm Res.* 2012;29(8):2202–12.
2. Philo JS. Characterizing the aggregation and conformation of protein therapeutics. *Am Biotechnol Lab.* 2003;23:22–4.
3. Philo JS. A critical review of methods for size characterization of non-particulate protein aggregates. *Curr Pharm Biotechnol.* 2009;10(4):359–72.
4. Filipe V, Jiskoot W, Basmeh AH, Halim A, Schellekens H, Brinks V. Immunogenicity of different stressed IgG monoclonal antibody formulations in immune tolerant transgenic mice. *MAbs.* 2012;4(6):740–52.
5. Rosenberg AS. Effects of protein aggregates: an immunologic perspective. *AAPS J.* 2006;8(3):E501–7.
6. Shire SJ, Shahrokh Z, Liu J. Challenges in the development of high protein concentration formulations. *J Pharm Sci.* 2004;93(6):1390–402.
7. den Engelsman J, Garidel P, Smulders R, Koll H, Smith B, Bassarab S, *et al.* Strategies for the assessment of protein aggregates in pharmaceutical biotech product development. *Pharm Res.* 2011;28(4):920–33.
8. Carpenter JF, Randolph TW, Jiskoot W, Crommelin DJ, Middaugh CR, Winter G. Potential inaccurate quantitation and sizing of protein aggregates by size exclusion chromatography: essential need to use orthogonal methods to assure the quality of therapeutic protein products. *J Pharm Sci.* 2010;99(5):2200–8.
9. Gabrielson JP, Brader ML, Pekar AH, Mathis KB, Winter G, Carpenter JF, *et al.* Quantitation of aggregate levels in a recombinant humanized monoclonal antibody formulation by size-exclusion chromatography, asymmetrical flow field flow fractionation, and sedimentation velocity. *J Pharm Sci.* 2007;96(2):268–79.
10. Laue TM. Analytical ultracentrifugation. Current protocols in protein science. USA: Wiley; 2001.
11. Liu J, Andya JD, Shire SJ. A critical review of analytical ultracentrifugation and field flow fractionation methods for measuring protein aggregation. *AAPS J.* 2006;8(3):E580–9.
12. Schuster TMTJ. New revolutions in the evolution of analytical centrifugation. *Curr Opin Struct Biol.* 1996;6:650–8.
13. Stafford III WF. Boundary analysis in sedimentation transport experiments: a procedure for obtaining sedimentation coefficient

- distributions using the time derivative of the concentration profile. *Anal Biochem.* 1992;203(2):295–301.
14. Cohen R, Claverie JM. Sedimentation of generalized systems of interacting particles. II. Active enzyme centrifugation—theory and extensions of its validity range. *Biopolymers.* 1975;14(8):1701–16.
 15. Claverie JM, Dreux H, Cohen R. Sedimentation of generalized systems of interacting particles. I. Solution of systems of complete Lamm equations. *Biopolymers.* 1975;14(8):1685–700.
 16. Demeler B, Saber H. Determination of molecular parameters by fitting sedimentation data to finite-element solutions of the Lamm equation. *Biophys J.* 1998;74(1):444–54.
 17. Schuck P. Sedimentation analysis of noninteracting and self-associating solutes using numerical solutions to the Lamm equation. *Biophys J.* 1998;75(3):1503–12.
 18. Schuck P, Demeler B. Direct sedimentation analysis of interference optical data in analytical ultracentrifugation. *Biophys J.* 1999;76(4):2288–96.
 19. Van Holde KE, Weischet WO. Boundary analysis of sedimentation-velocity experiments with monodisperse and paucidisperse solutes. *Biopolymers.* 1978;17(6):1387–403.
 20. Schuck P. On the analysis of protein self-association by sedimentation velocity analytical ultracentrifugation. *Anal Biochem.* 2003;320(1):104–24.
 21. Lamm O. The theory and method of ultra centrifuging. *Z Phys Chem A-Chem T.* 1929;143:177–90.
 22. Brown PH, Balbo A, Schuck P. On the analysis of sedimentation velocity in the study of protein complexes. *Eur Biophys J.* 2009;38(8):1079–99.
 23. Ghirlando R, Balbo A, Piszczek G, Brown PH, Lewis MS, Brautigam CA, *et al.* Improving the thermal, radial, and temporal accuracy of the analytical ultracentrifuge through external references. *Anal Biochem.* 2013;440(1):81–95.
 24. Schuck P. SEDFIT v 14.4f. 2014.
 25. Schuck P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys J.* 2000;78(3):1606–19.
 26. Brown PH, Balbo A, Schuck P. A bayesian approach for quantifying trace amounts of antibody aggregates by sedimentation velocity analytical ultracentrifugation. *AAPS J.* 2008;10(3):481–93.
 27. Gabrielson JP, Arthur KK. Measuring low levels of protein aggregation by sedimentation velocity. *Methods.* 2011;54(1):83–91.
 28. Gabrielson JP, Arthur KK, Stoner MR, Winn BC, Kendrick BS, Razinkov V, *et al.* Precision of protein aggregation measurements by sedimentation velocity analytical ultracentrifugation in biopharmaceutical applications. *Anal Biochem.* 2010;396(2):231–41.
 29. Dam J, Schuck P. Calculating sedimentation coefficient distributions by direct modeling of sedimentation velocity concentration profiles. *Methods Enzymol.* 2004;384:185–212.
 30. Schuck P. SEDFIT Help Web 2014 [cited 2014]. Available from: http://www.analyticalultracentrifugation.com/sedfit_help.htm.
 31. Schuck P, Perugini MA, Gonzales NR, Howlett GJ, Schubert D. Size-distribution analysis of proteins by analytical ultracentrifugation: strategies and application to model systems. *Biophys J.* 2002;82(2):1096–111.
 32. Paul R, Graff-Meyer A, Stahlberg H, Lauer ME, Rufer AC, Beck H, *et al.* Structure and function of purified monoclonal antibody dimers induced by different stress conditions. *Pharm Res.* 2012;29(8):2047–59.
 33. Gabrielson JP, Randolph TW, Kendrick BS, Stoner MR. Sedimentation velocity analytical ultracentrifugation and SEDFIT/c(s): limits of quantitation for a monoclonal antibody system. *Anal Biochem.* 2007;361(1):24–30.
 34. Brown PH, Balbo A, Schuck P. Using prior knowledge in the determination of macromolecular size-distributions by analytical ultracentrifugation. *Biomacromolecules.* 2007;8(6):2011–24.
 35. Brown PH, Schuck P. Macromolecular size-and-shape distributions by sedimentation velocity analytical ultracentrifugation. *Biophys J.* 2006;90(12):4651–61.
 36. Gabrielson JP, Arthur KK, Kendrick BS, Randolph TW, Stoner MR. Common excipients impair detection of protein aggregates during sedimentation velocity analytical ultracentrifugation. *J Pharm Sci.* 2009;98(1):50–62.
 37. Pekar A, Sukumar M. Quantitation of aggregates in therapeutic proteins using sedimentation velocity analytical ultracentrifugation: practical considerations that affect precision and accuracy. *Anal Biochem.* 2007;367(2):225–37.